

Respondent-driven Sampling on Directed Networks

Xin Lu^{a,b,*}, Jens Malmros^c, Fredrik Liljeros^b, Tom Britton^c

^a*Department of Public Health Sciences, Karolinska Institute, Stockholm, Sweden*

^b*Department of Sociology, Stockholm University, Stockholm, Sweden*

^c*Department of Mathematics, Stockholm University, Stockholm, Sweden*

Abstract

Respondent-driven sampling (RDS) is a commonly used substitute for random sampling when studying hidden populations, such as injecting drug users or men who have sex with men, for which no sampling frame is known. The method is an extension of the snowball sample method and can, given that some assumptions are met, generate unbiased population estimates. One key assumption, not likely to be met, is that the acquaintance network in which the recruitment process takes place is undirected, meaning that all recruiters should have the potential to be recruited by the person they recruit. Here we investigate the potential bias of directedness by simulating RDS on real and artificial network structures. We show that directedness is likely to generate bias that cannot be compensated for unless the sampled individuals know how many that potentially may have recruited them (i.e. their indegree), which is unlikely in most situations. Based on one known parameter, we propose an estimator for RDS on directed networks when only outdegrees are observed.

By comparison of current RDS estimators' performances on networks with varying structures, we find that our new estimator, together with a recent estimator, which requires the population size as a known quantity, have relatively low level of estimate error and bias. Based on our new estimator, sensitivity analysis can be made by varying values of the known parameter to take uncertainty of network directedness and error in reporting degrees into account. Finally, we have developed a bootstrap procedure for the new estimator to construct confidence intervals.

Keywords: Respondent-driven Sampling, directed networks, HIV

1. Introduction

Hidden populations (hard-to-reach populations), such as injecting drug users (IDU), men who have sex with men (MSM), and sex workers (SW) and their sexual partners, are generally considered as critical actors in the HIV epidemic [1, 2, 3]. Consequently, obtaining population characteristics and risk behaviors of these populations are critical for developing efficient disease control strategies. However, the lack of sampling frames for such populations makes traditional estimation methods based on random samples practically useless. Other methods have been proposed for such situations, for example key informant sampling [4], targeted/location sampling [5] and snowball sampling [6].

A more recent method is Respondent Driven Sampling (RDS), which was proposed to overcome difficulties when sampling hidden populations [7, 8, 9]. The RDS method starts with an initial selection of respondents, which are called "seeds". Each seed is given a number of "coupons" – tickets for participation in the study – to distribute to friends and acquaintances within the population of interest. When interviewed (anonymously), a new respondent is in turn given coupons to distribute. Everyone is rewarded both for completing the interview, and for recruiting their peers into the study. If the recruitment chains are sufficiently long, the sample composition will stabilize and become independent of the seeds. Additionally, information about who recruits whom and each respondent's personal network size (degree) are recorded.

* Address for correspondence: Xin Lu, Department of Sociology, Stockholm University, SE-106 91, Stockholm, Sweden.
Email address: lu.xin@sociology.su.se (Xin Lu)

There are two significant improvements in RDS compared to other non-random methods that has been used when sampling hidden population: Firstly, it uses dual incentives to encourage respondents to recruit more peers into the study. Secondly, when several assumptions are fulfilled, unbiased estimates of population proportions can be obtained by utilizing the recruitment and degree information collected in the sample.

Suppose a RDS study is performed on a connected undirected network with the additional assumptions that:

- (i) sampling of peer recruitment is done with replacement;
- (ii) each participant recruits one new participant to the study; and
- (iii) participants recruit randomly from their neighbors.

Then, the sampling probability of an individual i will be proportional to its degree when the sample reaches equilibrium. The population fraction p_A having a certain property A (e.g. p_A could denote the fraction among intravenous drug-users that are HIV-positive) can then be estimated by the weighted proportion of the sample fraction as in Volz and Heckathorn [9]:

$$\hat{p}_A^{VH_{out}} = \frac{\sum_{i \in U \cap A} d_i^{-1}}{\sum_{i \in U} d_i^{-1}}, \quad (1)$$

where the sample population U has been divided into two disjoint subsets A and $B = A^C$ depending on the reported properties of respondents, and d_i denotes the degree of individual i in the sample.

Based on equating the number of crossrelations between subgroups of property A and B , Salganik and Heckathorn [10] proposed an another widely used estimator for p_A :

$$\hat{p}_A^{SH_{out}} = \frac{\hat{s}_{BA} \hat{D}_B}{\hat{s}_{AB} \hat{D}_A + \hat{s}_{BA} \hat{D}_B}, \quad (2)$$

where $\hat{D}_A = \frac{n_A}{\sum_{i \in U \cap A} d_i^{-1}}$ and $\hat{D}_B = \frac{n_B}{\sum_{i \in U \cap B} d_i^{-1}}$ are the estimated harmonic mean degrees for the two subgroups. (n_A and n_B denote the number of A - and B -individuals in the sample respectively), and \hat{s}_{BA} denotes the sample fraction of all B -respondents naming A -peers and similarly \hat{s}_{AB} denotes the sample fraction of all A -respondents naming B -peers. For simplicity, we henceforth refer to (1) and (2) the VH_{out} and SH_{out} estimators respectively; the subscript *out* indicates that respondents out-degrees have been recorded, which will be important when we move to directed networks later on.

The ability to produce unbiased population estimates and a feasible field implementation have contributed to a rapid increase in RDS studies conducted globally in recent years [2, 11]. There has also been an increase in studies evaluating the performance of RDS estimators as well as in developing new estimators [12, 13, 14, 15].

Previous studies are mostly based on the assumption that relationships are reciprocal and the probability of being recruited by a peer is likewise equal on both ends of a relationship. Consequently, the network among which recruitments could take place is undirected. However, it is well-known that social networks, such as friendship networks, are generally directed to various extents. For example, in the study of Scott and Dana [16], only 6,669 out of 12,931 “best friend” nominations were found to be reciprocal, and in the study conducted by Wallace [17, 18], an average of 55.0 reciprocal nominations per respondent were found while the mean degree was 94.8. Evidence of irreciprocal recruiter-recruitee relationships has also been found in many RDS studies, e.g., in a RDS study of IDUs in Sydney, Australia [19], 29% of the respondents consider the relationship to their recruiter to be “not very close”, and in a study of IDUs in Tijuana, Mexico [20], only 62% of the respondents consider their relationship with their recruiter as “friend”. Additionally, in a study for MSM in Beijing, China [21], 8.5% participants said they received their coupons from a stranger, and between 3% to 7% of recruitments were found to be from strangers in the RDS studies on drug users and MSMs in three US cities and in St. Petersburg, Russia [22].

In [23], it has been shown that current RDS estimators may generate relatively large biases and errors if the studied networks are directed, indicating that estimates from previous RDS studies should be interpreted and generalized with caution. This study aims to further evaluate the influence of structural network properties on the performance of RDS estimators under the assumption that the underlying social network is (partially) directed, and to derive new

estimators allowing networks to be directed. In our evaluation, we use simulated data, as well as a real online MSM social network, to generate networks with varying directedness, degree correlation, indegree-outdegree correlation and homophily. Based on one known parameter, we propose a new estimator for RDS on directed networks and show how sensitivity analysis can be used to generate estimate intervals induced by intervals of unobserved network properties. Additionally, we develop a bootstrap procedure for the new estimator to construct confidence intervals.

2. RDS estimation on directed networks

We now investigate the properties of the RDS process on a directed network. For the purpose of this study, we focus on the problem of estimating the community fraction p_A having a certain dichotomous property A . Let G denote a (partially) directed network and let $e_{ij} = 1$ if there is a directed edge from i to j and $e_{ij} = 0$ otherwise. A reciprocal edge between i and j is hence reflected by $e_{ij} = e_{ji} = 1$. We assume that G is strongly connected, i.e., there is a directed path between any pair of nodes – otherwise we of course would not be able to estimate p_A well since it may then be impossible to reach certain parts of the community with RDS. Finally, we let N denote the community size, most often an unknown quantity in hidden or hard-to-reach populations. In what follows, assumptions (i)-(iii) are assumed to be fulfilled in the RDS process.

2.1. Extension of VH_{out} estimator to directed networks

When a RDS process takes place on a strongly connected network G , the recruitment of new respondents are dependent only on the current respondent, since he will select a new respondent uniformly from his peers. Thus, RDS possesses the Markov property [24] and can be modeled as a Markov process with transition matrix $R = \{a_{ij} = e_{ij}/d_i^{out}, 1 \leq i, j \leq N\}$, where d_i^{out} is the outdegree of node i [9, 23]. This process has a unique equilibrium distribution $\pi = [\pi_1 \cdots \pi_N]$ satisfying $R^T \pi^T = \pi^T$, indicating that π is the eigenvector corresponding to eigenvalue 1 for R^T . Consequently, π_i can be used to obtain the Hansen-Hurwitz estimator where observations are weighted by the inverse of the sampling probability [23]:

$$\hat{p}_A^{Eig} = \frac{\sum_{i \in U \cap A} \pi_i^{-1}}{\sum_{j \in U} \pi_j^{-1}}. \quad (3)$$

It has been shown that when the network is undirected, $\pi_i = d_i / \sum_{j=1}^N d_j$ is the analytical solution for π and \hat{p}_A can be estimated by the VH_{out} estimator: $\hat{p}_A^{VH_{out}} = \sum_{i \in U \cap A} d_i^{-1} / \sum_{i \in U} d_i^{-1}$.

Unfortunately, no analytical solution for π is available for a general directed network. However, note that under the above assumptions, the RDS process is merely a random walk on the network, for which we can easily adopt the mean field approach in [25] to derive an approximation of π :

Let $K \equiv (K_{in}, K_{out})$ be the set of nodes in the network with indegree K_{in} and outdegree K_{out} , and let f_K be the proportion of K -nodes in the set; then, the average inclusion probability of nodes in K is

$$\bar{\pi}(K) \equiv \frac{1}{N f_K} \sum_{i \in K} \pi_i. \quad (4)$$

Using that we have a random walk assumed to be in equilibrium, and taking the average over all nodes of degree K , we get

$$\bar{\pi}(K) = \frac{1}{N f_K} \sum_{i \in K} \sum_{j: k_{out}(j) \neq 0} \frac{e_{ji}}{k_{out}(j)} \pi_j, \quad (5)$$

where $k_{out}(j)$ is the outdegree of node j .

Then, the sum over j is parted into two, one over the degree classes K' and the other over the nodes within each degree class K' . We substitute π_j with the mean value within its degree class K' , yielding

$$\bar{\pi}(K) \simeq \frac{1}{N f_K} \sum_{K'} \frac{\bar{\pi}(K')}{K'_{out}} \sum_{i \in K} \sum_{j \in K'} e_{ji} = \frac{1}{N f_K} \sum_{K'} \frac{\bar{\pi}(K')}{K'_{out}} E_{K' \rightarrow K}, \quad (6)$$

where $E_{K' \rightarrow K}$ is the total number of edges pointing from nodes of degree K' to nodes of degree K , which we can write as $E_{K' \rightarrow K} = K_{in} f_K N \frac{E_{K' \rightarrow K}}{K_{in} f_K N} = K_{in} f_K N f_{K'|K}$, where $f_{K'|K}$ is the proportion of edges pointing to nodes in K originating in K' .

We finally obtain

$$\bar{\pi}(K) = K_{in} \sum_{K'} \frac{f_{K'|K}}{K'_{out}} \bar{\pi}(K'). \quad (7)$$

Particularly, when there is no degree dependence in the network, (7) becomes

$$\bar{\pi}(K) = K_{in} \sum_{K'} \frac{K'_{out} f_{K'|K} / \bar{K}_{in}}{K'_{out}} \bar{\pi}(K') = \frac{1}{N} \frac{K_{in}}{\bar{K}_{in}}, \quad (8)$$

where \bar{K}_{in} is the average indegree in the network, implying that for networks with no degree-degree correlations, the RDS sample can be weighted by respondents' indegrees to estimate population proportions, which gives us the modified VH_{out} estimator:

$$\hat{p}_A^{VH_{in}} = \frac{\sum_{i \in U \cap A} (d_i^{in})^{-1}}{\sum_{j \in U} (d_j^{in})^{-1}}. \quad (9)$$

Clearly, the difference between VH_{out} and VH_{in} estimators is the substitution of outdegree with indegree. Thus, the use of VH_{in} brings new challenges in the implementation of RDS as it requires collection of respondents' indegrees, which are not known from the RDS sample.

2.2. Extension of SH_{out} estimator to directed networks

The SH_{out} estimator was developed based on the fact that in any undirected network, the number of crossgroup edges pointing from A to B , equals the number of edges pointing from B to A . Similarly, in a directed network, the sum of nodes' indegrees in a group equals the total number of edges pointing to nodes in that group, i.e., if we let $S = \begin{bmatrix} S_{AA} & S_{AB} \\ S_{BA} & S_{BB} \end{bmatrix}$ be the recruitment matrix in the network, where, e.g., S_{AB} is the proportion of edges originating in group A which end in group B , then we have

$$\begin{cases} N_A \bar{D}_A^{out} S_{AA} + N_B \bar{D}_B^{out} S_{BA} = N_A \bar{D}_A^{in} \\ N_A \bar{D}_A^{out} S_{AB} + N_B \bar{D}_B^{out} S_{BB} = N_B \bar{D}_B^{in} \end{cases} \quad (10)$$

where, e.g., \bar{D}_A^{out} is the average outdegree in group A .

For simplicity, let $m^* = \frac{\bar{D}_A^{in}}{\bar{D}_B^{in}}$ and $w^* = \frac{\bar{D}_A^{out}}{\bar{D}_B^{out}}$ be the average indegree and outdegree ratio of the two groups of nodes in the network, and let $\phi = \frac{N_A}{N_B}$ be the relative group size proportion. Dividing the above equations (10) gives a solution of ϕ :

$$\phi = \frac{w^* S_{AA} - m^* S_{BB}}{2m^* w^* S_{AB}} + \sqrt{\frac{S_{BA}}{m^* w^* S_{AB}} + \left(\frac{m^* S_{BB} - w^* S_{AA}}{2m^* w^* S_{AB}} \right)^2}. \quad (11)$$

Then, if we can correctly estimate m^* , w^* and S , we obtain a generalization of the SH_{out} estimator:

$$\hat{p}_A^{SH_{in}} = \frac{\hat{\phi}}{1 + \hat{\phi}}, \quad (12)$$

in which we replace unknown population quantities in ϕ by their estimates from the RDS sample.

From the previous section, the average indegree ratio m^* in SH_{in} can be estimated by the harmonic mean ratio of indegrees from the sample for networks with no degree correlation: $\hat{m}^* = \frac{n_A / \sum_{i \in U \cap A} (d_i^{in})^{-1}}{n_B / \sum_{i \in U \cap B} (d_i^{in})^{-1}}$. It is however generally

not possible to consistently estimate w^* and S using only the average outdegree and observed recruitment matrix. The sample mean outdegree will be an unbiased estimator only if there is no correlation between the indegree and outdegree of nodes, while the harmonic mean of outdegree is expected to have higher precision if the indegree-outdegree correlation is high. However, in simulations it is seen that there is little difference in using either the (arithmetic) mean or harmonic mean of outdegree to estimate w^* and thereby we continue to use the harmonic mean in the following analysis. We have also tried to adjust potential bias in the estimation of S by replacing individual inclusion probabilities with group inclusion probabilities (see Supportive Information (SI) for details), which however didn't improve the results and we therefore prefer to use the observed recruitment matrix from the sample to estimate S in SH_{in} .

The factor w^* was named the *activity ratio* in [13], since it quantifies how active nodes in different groups are in building their personal networks. Following this, we henceforth refer to m^* as the *attractivity ratio*, as it reflects how "attractive" nodes in different groups are, or to which group of nodes edges are inclined to connect to.

2.3. Sensitivity analysis when indegree is not known

Hardly ever is the indegree observed in RDS studies. Consequently, the use of VH_{in} and SH_{in} is limited in practice. It is however possible to use both estimators if prior information is available. In SH_{in} , if the indegree is not known, the estimate of average indegree ratio, \hat{m}^* , becomes an unknown parameter in (12). This is true also for VH_{in} , since we can rewrite (9) as:

$$\hat{p}_A^{VH_{in}} = \frac{\sum_{i \in U \cap A} (d_i^{in})^{-1}}{\sum_{j \in U} (d_j^{in})^{-1}} = \frac{n_A / \hat{D}_A^{in}}{n_A / \hat{D}_A^{in} + n_B / \hat{D}_B^{in}} = \frac{n_A / n_B}{n_A / n_B + \hat{D}_A^{in} / \hat{D}_B^{in}}.$$

Replacing $\hat{D}_A^{in} / \hat{D}_B^{in}$ with m , we have:

$$\hat{p}_A^{VH_m} = \frac{n_A / n_B}{n_A / n_B + m}. \quad (13)$$

Prior information may, for example, be obtained by expert opinion, or by using previous empirical results. What's more, even if there is little prior knowledge about the targeted population, we can, instead of providing a point estimate with fixed parameters, use a range of m values to generate an estimate interval for p_A^* . That is, if m^* is assumed to lie within a certain range, $[m_{min}, m_{max}]$, we get an interval of \hat{p}_A , $[\hat{p}_A(m_{min}), \hat{p}_A(m_{max})]$, by varying m in (12).

Following this, we will perform a sensitivity analysis on SH_m , and VH_m , by varying the ratio of average indegrees m to get an interval of the estimates of p_A^* , $[\hat{p}_A(m_{min}), \hat{p}_A(m_{max})]$, with m lying in a certain range, $[m_{min}, m_{max}]$. By choosing an interval centered on a value of m based on prior information, we will get intervals of possible \hat{p}_A values which more fully accounts for the situation when the network is directed, and provides valuable results on the sensitivity of estimators to the correctness of indegree assumptions about the network.

3. Network Data and Study Design

We will evaluate the performance of our suggested estimators and compare them with existing estimators through simulations of RDS processes on directed networks. The simulations will be performed on both artificially generated families of directed networks as well as a real MSM online social network [23], which makes it possible to study the impact of different, carefully controlled, structural network properties on our estimators as well as looking at their behavior in a more realistic setting using actual data.

In our evaluation, we will consider the following parameters which are important both to directed networks and RDS estimation:

Directedness; if E_{dir} is the number of directed edges in a network with E edges, then the proportion of directed edges is:

$$\lambda = E_{dir} / E, \quad (14)$$

i.e., $\lambda = 0$ when the network is undirected, and $\lambda = 1$ when the network is (extremely) directed in a way such that there are no reciprocal edges.

Indegree correlation; the tendency that nodes with high indegrees are connected with each other. To quantify this, we use the assortativity ratio defined in [26]:

$$\gamma = \frac{E^{-1} \sum_i j_i k_i - [E^{-1} \sum_i \frac{1}{2}(j_i + k_i)]^2}{E^{-1} \sum_i \frac{1}{2}(j_i^2 + k_i^2) - [E^{-1} \sum_i \frac{1}{2}(j_i + k_i)]^2}, \quad (15)$$

where j_i and k_i are the indegrees of vertices at the end of the i^{th} edge, $i = 1, \dots, E$.

Indegree-outdegree correlation; unlike the indegree correlation, which describes associations between nodes, the indegree-outdegree correlation measures the correlation between indegree and outdegree for the same node. We use the Pearson correlation calculated from all nodes in the network:

$$\rho = \text{Cov}(d^{in}, d^{out}) / \sigma_{d^{in}} \sigma_{d^{out}}. \quad (16)$$

Homophily; the probability that nodes connect with neighbors that are similar to themselves with respect to the studied feature A rather than that they connect randomly [27, 28, 29, 8]. Letting h_A be the homophily for nodes with trait A , it holds that $S_{AA} = h_A + (1 - h_A)p_A$, implying that h_A can be calculated as:

$$h_A = 1 - S_{AB}/p_B. \quad (17)$$

The activity ratio w^* , as well as the attractivity ratio m^* , are also used as network structure parameters in our assessment.

3.1. Network Data

We will focus our study of network properties on directedness and attractivity ratio, as they are of general interest to the study of directed networks, and of particular interest to our estimators. Furthermore, we will vary other network structural properties that are likely to affect our estimators. For example, the VH_{in} estimator is based on the assumption of no indegree correlation in the network, and the estimate of average outdegree in SH_{in} is based on the assumption of positive indegree-outdegree correlation, etc.

In order to study the behavior of our estimators with respect to variation in directedness and attractivity ratio, we will use two families of generated networks, **Net1**, where there is little or no indegree correlation and no indegree-outdegree correlation, and **Net2**, which have varying homophily and positive indegree-outdegree correlation. This setting makes it possible to see how different structural properties, e.g. homophily, will affect our estimators as directedness and attractivity ratio are varied.

Net1 is generated starting from a random pure directed network, in which indegree and outdegree are uncorrelated ($\rho \approx 0$). Then, the irreciprocal edges are rewired in a particular way that doesn't change nodes' degree in order to generate networks with different levels of directedness (down to $\lambda = 0.2$) while the indegree-outdegree correlation remains unchanged. Finally, nodes are assigned either property A or B to achieve different attractivity ratios $m^* \in [0.7, 1.4]$ (see Table 1). The generating process for Net2 starts with a random undirected network. To obtain directedness, reciprocal edges are randomly rewired in such a way, that for any network in Net2 with directedness λ , the indegree-outdegree correlation is $\rho \approx 1 - \lambda$. Then, different attractivity ratios are generated as for Net1, and we further rewire edges with respect to nodes' properties in order to achieve different levels of homophily: $h_A \in [0, 0.5]$ (see Table 1). As we in this study are mostly interested in the case when sample size is relatively small compared to the population size, these networks are both of size 10,000.

Table 1: Basic statistics of Net1, Net2, Net3 and the MSM network

	Network size (N)	Average degree (\bar{D})	Directedness (λ)	indegree correlation (γ)	indegree-outdegree correlation (ρ)	Homophily (h)	Attractivity ratio (m^*)	P
Net1	10,000	10	[0, 1]	[-0.09, 0.01]	≈ 0	[-0.30, 0.22]	[0.7, 1.4]	70%
Net2	10,000	10	[0, 1]	[-0.03, 0.14]	$\approx 1 - \lambda$	[0, 0.5]	[0.7, 1.4]	30%
MSM Network	16,082	17.2	0.61	0.03	0.39	<i>age</i>	0.23	0.95
						<i>ct</i>	0.50	1.32
						<i>cs</i>	0.03	0.96
						<i>pf</i>	0.06	1.05
Net3	--*	--	[0.61, 0.91]	[0, 0.4]	--	--	--	--

* Same as the MSM network

The anonymized online social **MSM network** used in this study (previously analyzed in [30, 23]) has 16,082 nodes and comes from the Nordic region’s largest and most active web community for homosexual, bisexual, transgender and queer persons (www.qruiser.com) and includes information on the relationships between members as well as members’ personal information. Contacts between members on the web site are maintained by a “favorites list”, on which each member can add any other member without approval from that member, so that the resulting social network will be directed. From this network, we obtain the giant strongly connected component of the friendship network of those members who identify themselves as homosexual males. Utilizing information from their user profiles, we can evaluate our estimators on different personal characteristics, and we will focus on four dichotomous properties: age (born before 1980), county (live in Stockholm, ct), civil status (married, cs), and profession (employed, pf). The proportions of nodes having a specific value of these properties are listed in Table 1.

While this network provides an opportunity to study our estimators in a more realistic setting, it and the studied personal properties will obviously have certain structural properties. In order to keep this level of realism, while still varying some structural network properties, based on the MSM network, we generate a family of networks, **Net3**, which have different levels of indegree correlation ($\gamma \in [0, 0.4]$). Detailed information on the generation process of Net3 and the other networks can be found in the supportive information SI.

3.2. Simulation Design

RDS processes are then simulated on the above networks and estimates of RDS estimators are compared with true population properties. In each simulation, seeds are uniformly selected and coupons are randomly distributed to the recruiters’ neighbors. To simulate RDS in real practice, we let the number of seeds be 10 and the number of distributed coupons be 3 when shorter sample waves are desirable, and, 6 and 2 for longer sample waves (provided in the supporting information). Sampling is done without replacement and we choose sample size 500 for Net1 and Net2, and 1000 for the MSM network and Net3. All simulations are repeated 1000 times.

For each simulation, we estimate the population proportion with our suggested estimators as well as existing estimators. Then, the root mean square error (RMSE), standard deviation (SD) and bias of estimators are calculated in order to quantify the results. The estimators are divided into four categories:

- (i) The naïve estimator: The raw sample composition;
- (ii) Outdegree-based estimators: SH_{out} and VH_{out} ;
- (iii) Indegree-based estimators: SH_{in} and VH_{in} ;
- (iv) Estimators based on known parameters: SH_{m^*} and VH_{m^*} .

Note that the indegree-based estimators are practically useless, since individual indegrees are not known from the RDS sample, and are therefore presented merely for comparison and theoretical purposes.

Additionally, we include the estimator SS (Successive Sampling) suggested in [15] for reference; note that this estimator is based on a different estimation procedure and requires knowledge of the population size (N) in order to yield correct estimates. Since the SS estimator is developed for RDS on undirected networks, two versions of it will be used in order to adapt it for use with directed networks: SS_{out} using outdegrees of respondents in the sample in the estimation procedure and SS_{in} using their indegrees. In the estimation procedure of SS_{out} and SS_{in} , $M = 500$ times successive sampling samples per each of $r = 3$ iterations are used.

4. Results

4.1. Estimation performance on networks with known properties

4.1.1. Networks with varying structural properties

We start by looking at how varying directedness and attractivity ratio affects RDS estimators in an otherwise uncomplicated setting, i.e., the generated networks with close to zero indegree correlation and no indegree-outdegree correlation, **Net1**. In Figure 1, the bias, SD and RMSE of the raw sample composition, VH_{out} and VH_{m^*} are shown in the top row (SH_{in} , SH_{m^*} , and VH_{in} perform very similar to VH_{m^*} and are thus left out), and the same figures for VH_{m^*} , SS_{out} , and SS_{in} are shown in the bottom row for visual clarity.

We can see in the top row that both the raw sample composition and VH_{out} are biased with increasing $|m^* - 1|$, and that VH_{out} has the same level of bias and RMSE as the sample composition as long as the network is directed,

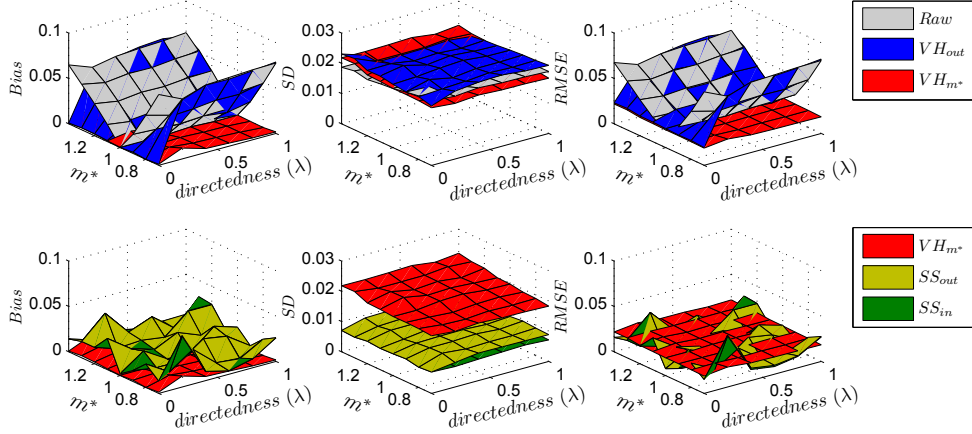


Figure 1: Bias, Standard Deviation, and Root Mean Square Error of RDS estimators on Net1. Sampling without replacement, number of seeds=10, coupons=3, sample size=500.

i.e., $\lambda > 0$. On the other side, the indegree-based estimator, VH_{m^*} , generates negligible bias and consistently smaller RMSE.

In the bottom figures, we see that SS_{in} and SS_{out} have varying bias (smaller than VH_{out}) as directedness and attractivity ratio changes, while retaining a substantially lower SD; on the other hand, VH_{m^*} has smaller bias but larger SD. Consequently, the RMSE of SS_{in} and SS_{out} becomes similar with that of VH_{m^*} due to their small SD; sometimes, the RMSE of SS_{in} and SS_{out} is even smaller than that of VH_{m^*} .

The results in Figure 2 for RDS on networks with varying indegree-outdegree correlation, but no homophily, **Net2**, are similar to those seen in Figure 1, except that the bias and RMSE of VH_{out} now increase gradually with the directedness of the network, generating bias and RMSE smaller than the raw sample composition, but larger than VH_{m^*} . This difference in performance of VH_{out} on Net1 and Net2 shows that for RDS on network with indegree-outdegree correlation, the traditional outdegree-based estimators, can still be expected to give less estimate bias and error than the raw sample composition. However, the indegree-outdegree correlation have little effect on the performance of estimators utilizing known parameters, i.e., SS_{in} , SS_{out} , and VH_{m^*} .

In Figure 3, where homophily $h = 0.4$, we see that the magnitude of bias, SD and RMSE all increase for the raw sample composition, VH_{out} and VH_{m^*} , indicating a clear effect of homophily on increasing RDS estimate bias and error. However, on the other hand, SS_{in} and SS_{out} are quit robust to the effect of homophily, their SD ([0.008, 0.01]) remains substantially smaller than the rest estimators ([0.02, 0.04]), and in most cases they produce minimum RMSE.

Overall, it is clear that previous RDS estimators are seriously affected by letting RDS processes take place on directed networks, and that our suggested estimators and the SS estimators, although all relying on previous knowledge about the network, shows major improvements in the quality of estimates.

4.1.2. The MSM network and its modifications

In the **MSM network** we look at four dichotomous user properties and how the estimators behave on each of them. As the structural properties of the MSM network are fixed, we illustrate estimator behavior through box plots, which are shown in the left panel of Figure 4. In each box, the central line is the median, the dot is the mean, the edges of the box are the 25th (q_1) and 75th (q_3) percentiles. Estimates being at least $1.5(q_3 - q_1)$ away from the edges of the box are shown as outliers beyond the whiskers.

The traditional outdegree-based estimators, SH_{out} and VH_{out} , have large bias when estimating variables with large homophily and attractivity ratios which significantly differ from 1, i.e., age and county. For example, their estimates of the proportion of MSM members who live in Stockholm are on average over 5 percentage units higher than the true value, and for age, civil status and profession, the sample mean has even less bias than them.

The indegree-based estimators, SH_{in} and VH_{in} , are generally much less biased for all variables, indicating that the indegree is a good approximation of sampling probability for nodes in directed networks. The m^* -based estimators,

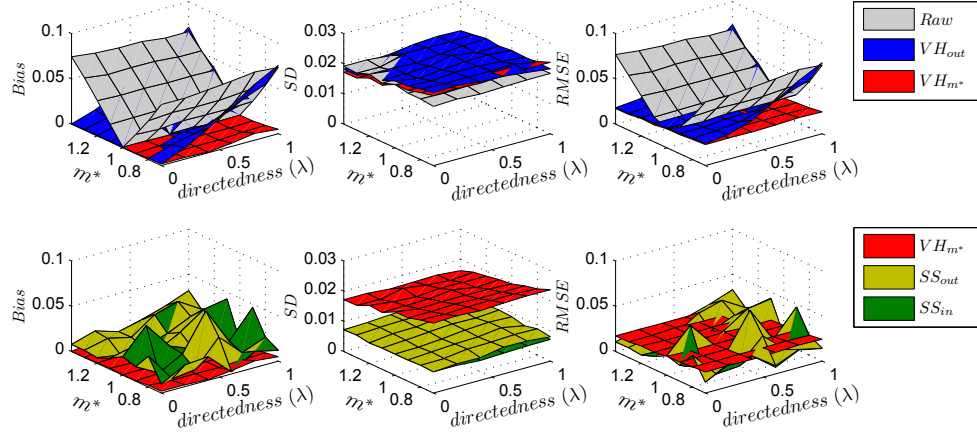


Figure 2: Bias, Standard Deviation, and Root Mean Square Error of RDS estimators on Net2, homophily $h_A = 0$. Sampling without replacement, number of seeds=10, coupons=3, sample size=500.

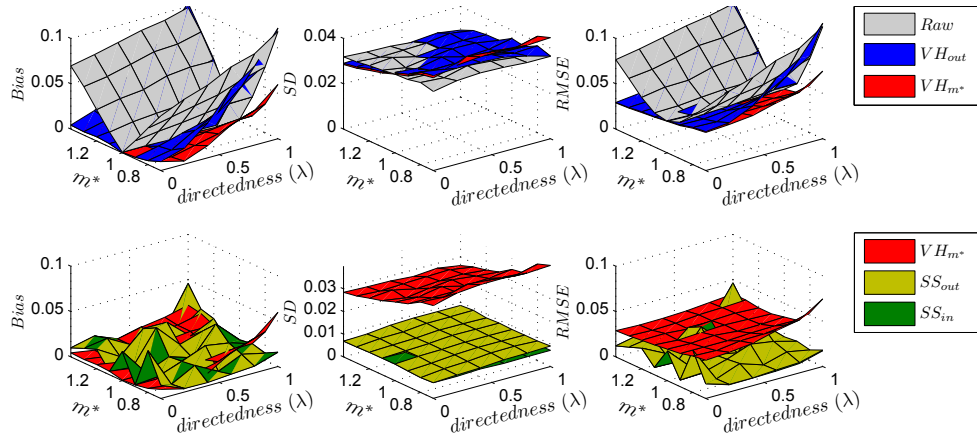


Figure 3: Bias, Standard Deviation, and Root Mean Square Error of RDS estimators on Net2, homophily $h_A = 0.4$. Sampling without replacement, number of seeds=10, coupons=3, sample size=500.

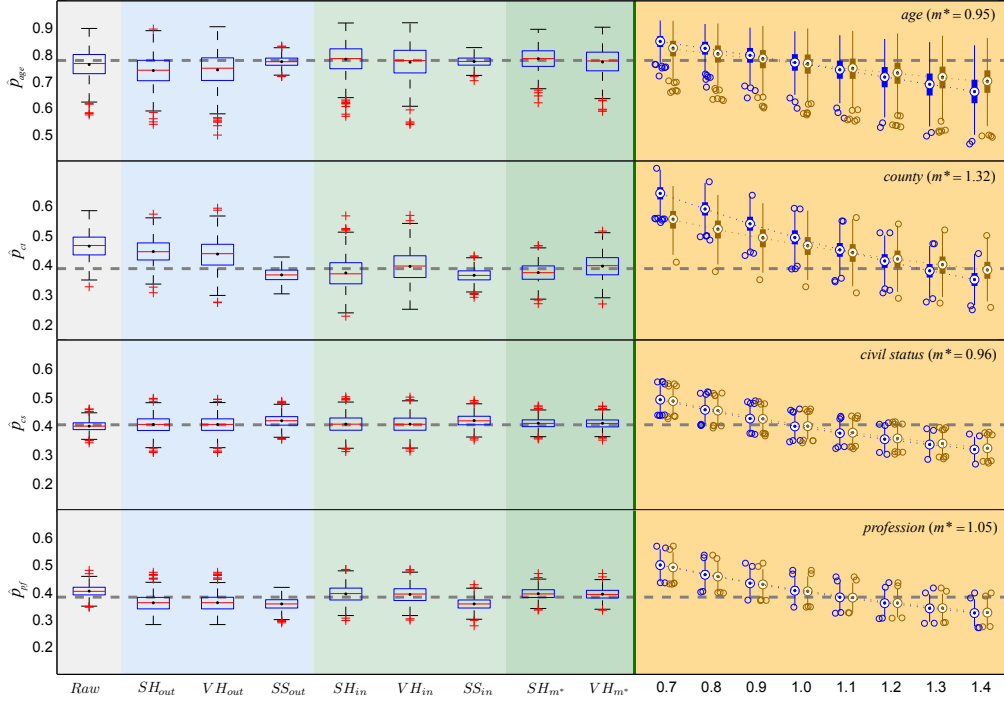


Figure 4: RDS on MSM network. The right panel shows sensitivity analysis of $\hat{p}_A^{VH_m}$ (brown) and $\hat{p}_A^{SH_m}$ (blue) with m varying from 0.7 to 1.4, plots are horizontally shifted a few points to avoid overlapping. Sampling with replacement, number of seeds=10, number of coupons=3, sample size=1000.

SH_{m^*} and VH_{m^*} , have a similar performance on the biasedness, with slightly smaller SD.

Lastly, when we look at the N -based estimators, SS_{out} and SS_{in} , they are biased for all the four variables, however, the (substantially) smaller SD makes their error lie within an acceptable range compared to SH_{out} and VH_{out} .

In Figure 5, we can see that the results from simulations on the modified MSM network, **Net3** with indegree correlation $\gamma = 0.4$, are very similar to the results from the unmodified MSM network. The indegree correlation gives the indegree-based estimators, SH_{in} and VH_{in} , together with the m^* -based estimators, SH_{m^*} and VH_{m^*} , a slightly increased bias, however, the overall performance of these estimators are better than SH_{out} and VH_{out} .

The results from the two SS estimators are practically unchanged from the MSM network, which indicates that these estimators are very robust to changes in indegree correlation. Again we find (substantially) smaller SD over all variables generated by these estimators, which makes the estimated error of SS_{out} and SS_{in} comparatively small despite that they are biased on directed networks.

4.2. Sensitivity analysis

We perform sensitivity analysis on SH_m and VH_m with respect to the attractivity ratio m . The results from the generated networks can be seen in Figure 6.

Figure 6(a) shows how the RMSE of VH_m changes with directedness and attractivity ratio when different values of m are given to the estimator as simulations are performed on **Net1**. It is clear that proper prior information on m will give small error in the estimator; note also that changes in directedness does not affect estimator performance.

Figures 6(b) and 6(c) shows the RMSE of VH_m and SH_m from simulations on **Net2** with homophily $h = 0$ and $h = 0.4$ respectively, and it can be seen that while the estimators have similar performance when homophily is low, VH_m generate less RMSE when m is far away from m^* and homophily is high, implying that when m^* is not known, VH_m may be a better option than SH_m in real practice.

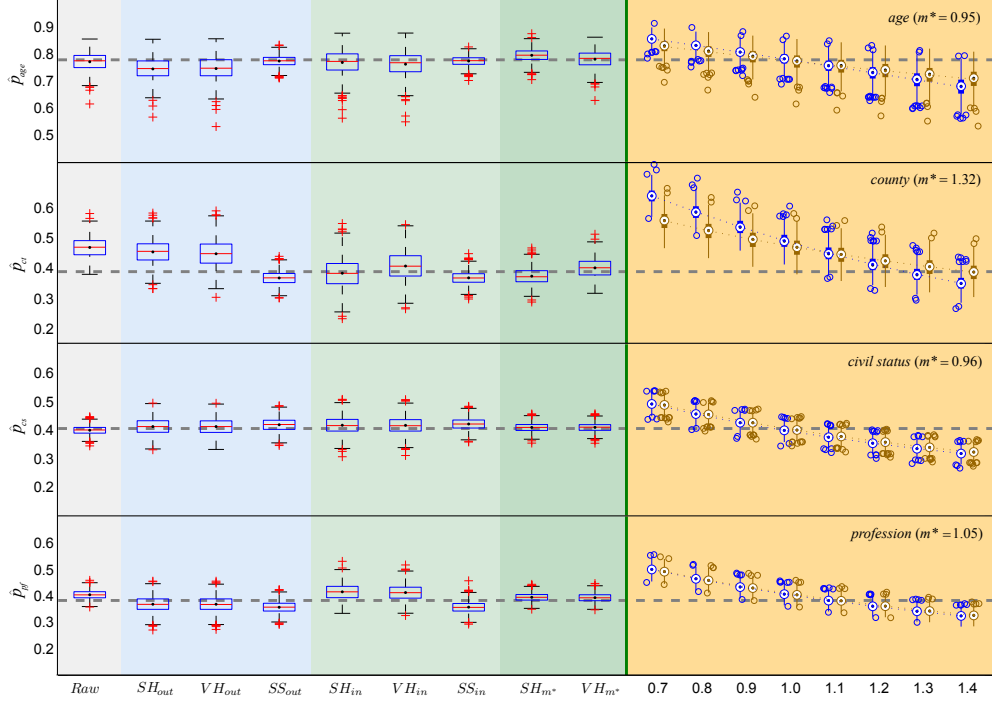


Figure 5: RDS on Net3 with indegree correlation $\gamma = 0.4$. The right panel shows sensitivity analysis of $\hat{p}_A^{VH_m}$ (brown) and $\hat{p}_A^{SH_m}$ (blue) with m varying from 0.7 to 1.4, plots are horizontally shifted a few points to avoid overlapping. Sampling without replacement, number of seeds=10, number of coupons=3, sample size=1000.

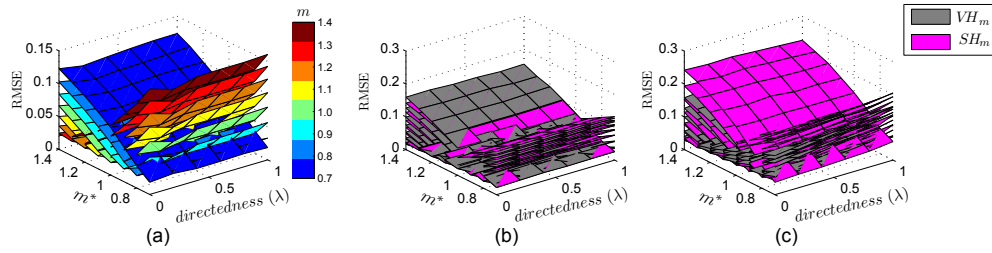


Figure 6: Sensitivity analysis of $\hat{p}_A^{VH_m}$ and $\hat{p}_A^{SH_m}$ on Net1 and Net2 with tested m values. (a) Net1, $\hat{p}_A^{SH_m}$ not shown as it is similar to $\hat{p}_A^{VH_m}$; (b) Net2 with homophily $h_A = 0$; (c) Net2 with homophily $h_A = 0.4$. Sampling without replacement, number of seeds=10, coupons=3, sample size=500.

In the sensitivity analysis on the **MSM network** and **Net3**, which can be seen on the right side of Figures 4 and 5, there is more variability in the estimates from age and county than from profession and civil status, as would be expected from the previous results. We see that the change in VH_m is smaller than in SH_m as m is varied; this is however negligible for profession and civil status. Generally, we see that we will cover the true value well by using the average estimates from the sensitivity analysis, which is especially interesting for county, the only property of which m^* significantly differs from 1.

Overall, we can see that VH_m performs better than SH_m , making it the preferred choice for RDS in real practice. We also did simulations on the above networks with 6 seeds and 2 coupons; however, no substantial differences are found in the results on the performance of estimators nor for the sensitivity analysis, see supplementary material.

4.3. Confidence interval and implementation

An problem associated with the use of VH_m , is to construct a confidence interval around $\hat{p}_A^{VH_m}$ when $m^* = m$. Traditionally, the standard error of RDS estimates are generated by a bootstrap procedure [31], in which replicated samples are drawn based on the recruitment property of original RDS samples. We modify the traditional bootstrap method by letting $\hat{p}_A^{VH_m}$ substitute the traditional RDS estimator when each bootstrapped sample is produced, and then let the middle 90% (95%) of the ordered estimates from the bootstrap samples' estimates be the approximation of the confidence interval.

We test the above procedure on Net1 and Net2; for each simulation setting $([\lambda, m^*, h_A])$, we take 1000 RDS samples and for each of these 1000 samples we construct 90% and 95% confidence intervals based on 1000 replicate samples drawn by the above bootstrap procedure. The proportion of times that the generated confidence interval contains the true population value p_A^* , denoted as Φ^{90} and Φ^{95} , are compared with the coverage rates of the traditional RDS estimator based method and are presented in Figure 7.

Apparently, due to the large bias of VH_{out} when network directedness and attractivity ratio is high, the traditional bootstrap procedure performs quite poorly with respect to Φ^{90} and Φ^{95} . The attractivity ratio has substantial impact; when $m^* = 1$, the coverage rate is generally close to the desired confidence level; however, the coverage rate drops rapidly as m^* deviates away from 1. For example, when $m^* = 0.8$, the coverage rate of the 90% confidence interval is only 29% even when the network (Net1) is undirected.

On the other hand, the VH_m -based bootstrap procedure gives reliable and consistent confidence intervals over all network settings. With the exception of some extreme cases, i.e., when $m^* \leq 0.9$, $\lambda > 0.6$, and $h_A = 0.4$, the coverage rate is fairly close to the desired confidence level and are overall better than that of the VH_{out} -based procedure.

It is of interest to use the previously suggested sensitivity analysis together with the given bootstrap procedure. As an illustration on how to implement the proposed methods in real RDS practice, when indegree information is not collected, we take data given in [32] and perform sensitivity analysis with VH_m , providing confidence intervals for all values of m . A sample of 618 drug users in New York City, and their personal characteristics, were collected using RDS with eight seeds. By using our methods on this data, we produce estimates and 90% confidence intervals on the proportion of males and the proportion of injectors.

It is not obvious which values of m that should be used in the sensitivity analysis. One suggestion is to let m vary around the observed activity ratio \hat{w}^* , since the indegree-outdegree correlation is positive in most social networks [16, 17, 18]. The activity ratio (\hat{w}^* , weighted) for males is 0.99, indicating that there is little difference of the size of personal networks with respect to gender. However, the activity ratio for injectors is 1.58, indicating that injecting drug users know substantially many more drug users than those who don't inject drugs. The length of the interval of m is arbitrarily set to 1.

In Fig. 8, we can see that when $m = \hat{w}^*$, the VH_m estimates are equal to those given by VH_{out} . When the network is assumed directed and $m \in [0.5, 1.5]$, the estimated proportion of male drug users will vary from 0.88 to 0.66. The proportion of injecting drug users, varies from 0.45 to 0.62 when $m \in [1, 2]$. The m intervals used here are arbitrarily chosen and their precision thus unknown, and therefore, it is hard to draw major conclusions from this example. However, the above analysis conveys another important information: for each change of 0.1 in the average indegree ratio, the change in the RDS estimates will be about 2 percentage units, which also may be an indication of how sensitive the RDS estimates are to uncertainties in the collected degree data.

	m^*							
	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
directedness (λ)								
0.0	.04 (.91)	.29 (.91)	.74 (.91)	.89 (.91)	.77 (.91)	.42 (.91)	.19 (.90)	.07 (.91)
0.2	.22 (.86)	.54 (.89)	.82 (.90)	.89 (.92)	.83 (.91)	.72 (.90)	.46 (.88)	.31 (.86)
0.4	.07 (.90)	.40 (.90)	.81 (.88)	.90 (.89)	.80 (.91)	.56 (.88)	.35 (.88)	.15 (.90)
0.6	.07 (.90)	.40 (.90)	.73 (.89)	.89 (.90)	.77 (.90)	.57 (.92)	.29 (.90)	.16 (.90)
0.8	.06 (.89)	.34 (.89)	.80 (.91)	.92 (.90)	.78 (.89)	.49 (.90)	.27 (.92)	.11 (.91)
1.0	.04 (.91)	.33 (.92)	.75 (.90)	.90 (.91)	.76 (.91)	.49 (.92)	.23 (.91)	.08 (.90)
Φ_{Net1}^{90}								
	m^*							
	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
directedness (λ)								
0.0	.16 (.91)	.42 (.90)	.74 (.91)	.88 (.90)	.75 (.92)	.37 (.91)	.10 (.91)	.01 (.90)
0.2	.17 (.90)	.46 (.91)	.77 (.90)	.91 (.91)	.77 (.91)	.42 (.90)	.12 (.91)	.02 (.90)
0.4	.15 (.91)	.44 (.91)	.75 (.91)	.90 (.91)	.74 (.89)	.43 (.91)	.12 (.90)	.03 (.90)
0.6	.13 (.90)	.45 (.90)	.72 (.90)	.90 (.90)	.77 (.91)	.46 (.91)	.20 (.92)	.05 (.90)
0.8	.09 (.88)	.38 (.90)	.74 (.91)	.90 (.90)	.80 (.91)	.48 (.89)	.20 (.91)	.06 (.88)
1.0	.06 (.87)	.35 (.92)	.71 (.91)	.91 (.91)	.80 (.90)	.46 (.90)	.21 (.89)	.06 (.88)
$\Phi_{\text{Net2}, h_A=0}^{90}$								
	m^*							
	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
directedness (λ)								
0.0	.24 (.91)	.55 (.89)	.84 (.91)	.94 (.90)	.84 (.90)	.49 (.90)	.17 (.92)	.02 (.91)
0.2	.25 (.88)	.58 (.89)	.86 (.89)	.95 (.91)	.86 (.90)	.54 (.91)	.20 (.90)	.04 (.91)
0.4	.23 (.80)	.58 (.87)	.84 (.89)	.96 (.90)	.84 (.92)	.57 (.90)	.21 (.90)	.07 (.91)
0.6	.19 (.62)	.56 (.84)	.82 (.89)	.95 (.90)	.86 (.91)	.60 (.88)	.31 (.87)	.09 (.84)
0.8	.14 (.49)	.50 (.76)	.84 (.84)	.95 (.89)	.88 (.91)	.61 (.85)	.30 (.84)	.11 (.82)
1.0	.10 (.26)	.47 (.60)	.82 (.86)	.95 (.92)	.88 (.90)	.60 (.86)	.30 (.82)	.11 (.79)
$\Phi_{\text{Net2}, h_A=0.4}^{90}$								
	m^*							
	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4
directedness (λ)								
0.0	.81 (.94)	.88 (.95)	.93 (.95)	.95 (.94)	.93 (.95)	.88 (.95)	.78 (.97)	.66 (.96)
0.2	.68 (.93)	.82 (.93)	.90 (.94)	.95 (.96)	.93 (.94)	.80 (.96)	.68 (.96)	.51 (.95)
0.4	.47 (.87)	.72 (.93)	.90 (.95)	.96 (.95)	.91 (.96)	.80 (.96)	.59 (.95)	.33 (.95)
0.6	.24 (.72)	.64 (.90)	.87 (.95)	.95 (.95)	.89 (.95)	.72 (.94)	.41 (.93)	.21 (.91)
0.8	.11 (.61)	.46 (.85)	.79 (.90)	.95 (.95)	.89 (.95)	.63 (.92)	.38 (.91)	.15 (.89)
1.0	.02 (.37)	.24 (.72)	.81 (.91)	.96 (.96)	.88 (.96)	.56 (.92)	.23 (.89)	.07 (.86)
$\Phi_{\text{Net2}, h_A=0.4}^{95}$								

Figure 7: 90% and 95% bootstrap coverage probability of $\hat{\rho}_A^{VH_{out}}$ and $\hat{\rho}_A^{VH_{m^*}}$ (shown in brackets) on Net1 and Net2. Sampling without replacement, number of seeds=10, coupons=3, sample size=500.

5. Conclusion and Discussion

Despite the widely acknowledged evidence of the existence of directedness among social networks, the effect of directedness on RDS estimates has seldom been evaluated. This could be problematic since all previously reported RDS estimates rely on the assumption that the studied networks are purely reciprocal, the violation of which will result in unknown bias. To address this situation, we have extended previous RDS estimators onto directed networks and evaluated their performance on networks with various structural properties.

Our study shows that the individual indegree is a fair approximation to nodes' sampling probabilities in a RDS process on a directed network, and that this approximation is robust to changes in indegree-outdegree correlation and indegree correlation etc. The suggested indegree-based estimators SH_m and VH_m can also be used with prior information on m^* , the ratio of harmonic mean indegrees of the two groups in the sample; an approach that gives good results. Prior information about m^* may possibly be obtained by expert opinions, or by using previous empirical studies related to the studied population; the need for prior information however limits this application in practice.

To further address the reality that the indegrees are not observed from an RDS sample, we have developed a sensitivity analysis method, based on the attractivity ratio m^* between the studied groups in the population, to incorporate the uncertainties in both network directedness and reported outdegrees. Our results show that, while it is of course best to have correct prior information on the network, it is possible to get a deeper understanding of how RDS estimation is influenced by network directedness by using sensitivity analysis. E.g., in our results from the MSM network it is shown that erroneous prior information may give serious errors.

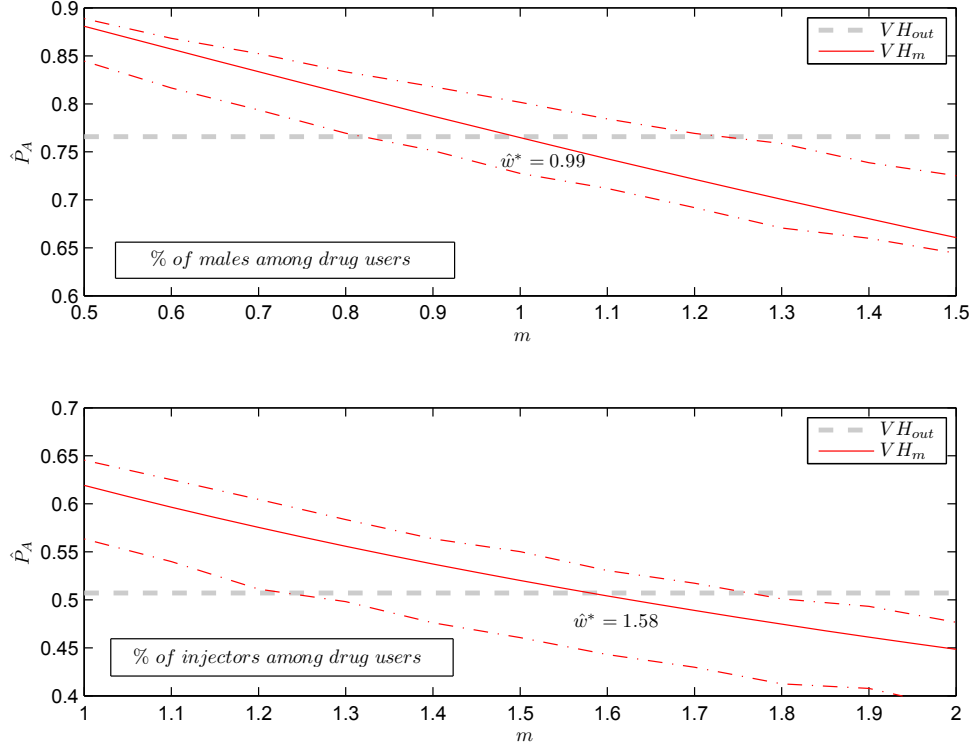


Figure 8: Sensitivity analysis of RDS estimates for proportion of (a) males and (b) injectors among drug users in New York City. The dash-dot line shows 90% CI of $\hat{p}_A^{VH_m}$.

It is difficult to find an obvious setup for the sensitivity analysis both with respect to guessing a value of m^* and deciding to which extent it should be varied. However, since many social networks have positive indegree-outdegree correlation, the activity ratio \hat{w}^* , which is observed from the sample, may be an indicator of where to vary m from, and often the difference between m^* and w^* is small. For example, in the MSM network, the absolute difference is 0.27, 0.17, 0.02, and 0.05 for age, county, civil status and profession, respectively [23].

From the results of sensitivity analysis on Net1 and Net2, we can see that the performance of VH_m and SH_m is determined primarily by the attractivity ratio m^* , rather than network directedness λ . Thus, if the network instead is assumed undirected, in which the ratio of indegrees is equal to the ratio of outdegrees ($m^* = w^*$), the sensitivity analysis may instead be used to assess the uncertainty of reported (out)degrees. The differential function of VH_m over m , $\frac{\partial VH_m}{\partial m} \big|_{m=\hat{w}^*} = \left(\frac{n_A}{n_B + m} \right)' \big|_{m=\hat{w}^*} = -\frac{n_A}{(n_B + \hat{w}^*)^2}$, then provides the magnitude of how much the RDS estimate would change if there is any reporting error in the degree information.

Acting as a natural companion to the sensitivity analysis or estimation with prior information, our modification of the previous RDS bootstrap method has proven to work well on directed networks. Based on the VH_m estimator, it largely outperforms the traditional VH_{out} -based method, and can construct close-to-expected confidence intervals for networks with varying directedness and attractivity ratio.

Another finding, which has not been highlighted in previous research, concerns the SS estimator [15]. This estimator has small and overall consistent standard error among the networks tested in our paper. Given that the population size is known, this estimator is expected to produce RDS estimates with acceptable bias and error.

While it is in the interest of this study to do a full evaluation of RDS on directed networks, it is worth noting that since RDS utilizes a peer-driven mechanism, and the recruitment rights are few and valuable, respondents are usually

inclined to recruit those that they know reasonably well. Such a mechanism to a large extent avoids the occurrence of recruitment via directed edges, and in most RDS studies, the proportion of recruitment through strangers are relatively small, usually less than 10% [33, 22, 21, 19, 20]; as pointed out earlier, this proportion may in some cases be larger though. In our results, we see that already small proportions of directed edges affects previous estimators, so while networks with extremely high directedness is very unlikely to occur in reality, and primarily are included out of theoretical interests, our evaluation shows that estimators will be sensitive to directedness in reality and that it is therefore an important issue to address.

For actual RDS practice, network directedness has previously not been an highlighted issue. The suggested sensitivity analysis gives RDS practitioners the possibility to take directedness into account as it provides means to understand the robustness of sample inference to the violation of certain assumptions: that the network may be partially directed, and that the degree data collected from respondents may contain reporting error. As there are no methods available on how to quantify network directedness, an interval of estimates based on a range of m values is currently the best way of understanding this issue; additionally, it may give researchers a more detailed image of the situation and advice on how to understand the studied population.

We hope that the current study can inspire research beyond the purpose of studying hidden populations, such as sampling the contents of webpages, where indegree may be used as a cheap and efficient parameter to approximate the inclusion probability of random walks on internet [25, 34, 35, 36]. Other factors that might affect inclusion probabilities, such as transitivity, degree distribution, closeness, etc., yet need to be investigated in future studies.

Acknowledgments

This work is funded in part by Riksbankens Jubileumsfond (dnr: P2008-0674) and the Swedish Research Council. X.L. would like to thank China Scholarship Council (Grant No. 2008611091). Thanks are also due to Sida for their support to RDS development work in Vietnam.

Appendix A: Generation process of Net1, Net2 and Net3

A.1. Net1

Net1 is the set of networks with different levels of directedness, in which the indegree and outdegree are not correlated ($\rho \approx 0$).

Step 1 Base network. At first, a random purely directed network ($\lambda = 1$) is generated by randomly distributing $N\bar{D}$ irreciprocal/directed edges between N nodes with the restriction that no reciprocal edges should be formed.

Step 2 Varying directedness. In order to decrease the directedness to a given $\lambda \in [0.2, 1]$, irreciprocal edges in the base network are randomly chosen and rewired to form reciprocal edges. Specifically, at each step, an edge $i \rightarrow j$ between nodes v_i and v_j is randomly chosen; then, if there is no link pointing from j to i , we randomly find an irreciprocal incoming edge of i , $k \rightarrow i$ and an irreciprocal outgoing edge of j , $j \rightarrow l$ (Figure 9(a)). These edges are then rewired as $k \rightarrow l$, and $j \rightarrow i$ (Figure 9(b)), such that a new reciprocal pair of edges $i \leftrightarrow j$ is formed, and the degrees of i , j , k and l remain unchanged. The rewiring process is restarted from the beginning if the network is disconnected.

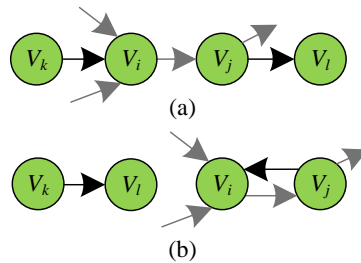


Figure 9: Net1: Illustration of the rewiring process leading to a decrease in λ

Step 3 Varying attractivity ratio. Let m^* be the desired attractivity ratio. For each network generated in **Step 2**, NP^* ($0 < P^* < 1$) nodes are randomly picked and assigned property A, and the remaining nodes are assigned property B. Then, the attractivity ratio of the network, $m' = \frac{\bar{d}_A^{in}}{\bar{d}_B^{in}}$, where \bar{d}_A^{in} , \bar{d}_B^{in} are the average indegrees of nodes with property A and B respectively, is calculated. If $m' \neq m^*$, the following algorithm is carried out in order to generate a network with the required m^* value:

- (i) Randomly pick two nodes, v_i and v_j , with different properties;
- (ii) If $m' > m^*$ and $\bar{d}_A^{in} > \bar{d}_B^{in}$, exchange the property of v_i and v_j ;
- (iii) Else if $m' < m^*$ and $\bar{d}_A^{in} < \bar{d}_B^{in}$, exchange the property of v_i and v_j ;
- (iv) Repeat (i)-(iii) until $m' = m^*$.

Step 4 A random undirected network of the same size and average degree is generated separately for $\lambda = 0$, and the method described in **Step 3** is used to generate different m^* values in this network.

A.2. Net2

Net2 is the set of networks with a certain amount of indegree-outdegree correlation and different levels of directedness and homophily.

Step 1 Base network. At first, a random purely undirected network ($\lambda = 0$) is generated by randomly distributing $N\bar{D}/2$ reciprocal/undirected pairs of edges between N nodes.

Step 2 Varying directedness. In order to increase the directedness to a given $\lambda \in [0, 1]$, reciprocal edges in the base network are randomly chosen and rewired to form irreciprocal edges. Specifically, in each step, a pair of reciprocal edges $i \leftrightarrow j$ is randomly chosen. If there is no link between two randomly chosen nodes k and l (Figure 10(a)), then we randomly pick one of the two links between i and j and use it to connect k and l (Figure 10(b)). Such a process leads to an indegree-outdegree correlation $\rho \approx 1 - \lambda$. The rewiring process is restarted from the beginning if the network is disconnected.

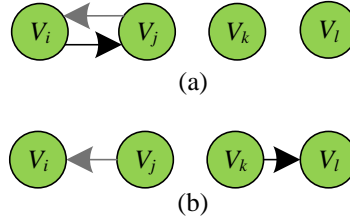


Figure 10: Net2: Illustration of the rewiring process carried out to increase λ

Step 3 Varying attractivity ratio. The same process as described in **Step 3** in Section 5 is used to generate different m^* values for each network generated in **Step 2**.

Step 4 Homophily. In order to generate networks with different homophily for group A, h_A , links are further rewired in each network generated in **Step 3**. Let h'_A be the homophily of group A in the current network. At each step, either a pair of irreciprocal links or reciprocal links are randomly picked (Figure 11); if $h'_A > h_A$, meaning that there are too many within-group connections, we rewire the within group links, $i \rightarrow k$, $l \rightarrow j$ (or $i \leftrightarrow k$, $l \leftrightarrow j$), to $i \rightarrow j$, $k \rightarrow l$ (or $i \leftrightarrow j$, $k \leftrightarrow l$), or vice versa if $h'_A < h_A$. The above process is repeated until $h'_A = h_A$.

A.3. Net3

Net3 is the set of networks with different levels of indegree correlation ($\gamma \in [0, 0.4]$), varied from the MSM network; all four node properties (*age*, *county*, *civil status* and *profession*) are kept.

Step1 Base network. The MSM friendship network obtained from the web community [30, 23].

Step2 Varying γ . A shuffling method slightly different from what was described in [37] is used to generate networks with different indegree correlation. At each step, we randomly pick a pair of edges, $i \rightarrow j$, $k \rightarrow l$ (Figure 12). If the indegrees of i and l are the two largest or the two smallest among the four nodes, and j and l have the same property, we rewire the two edges as $i \rightarrow l$, $k \rightarrow j$. Then, the degree distribution and homophily of the network is kept, and the indegree correlation increases as the rewiring process progresses. We generate networks with γ up to 0.4 for each of the four properties in the MSM network.

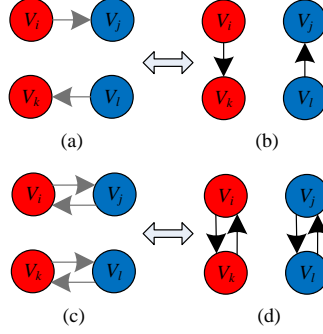


Figure 11: Net2: Illustration of the rewiring process resulting in a change of h_A . Red: trait A, Blue: trait B

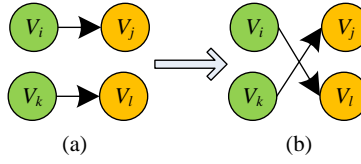


Figure 12: Net3: Illustration of the shuffling process used for network generation. V_j and V_l have the same property.

Appendix B: Discussions on the estimate of S_{XY} in SH_{in}

It can be proven that when the network is undirected, a node will be recruited into a RDS sample with a probability proportional to its degree if the assumptions for SH_{out} are fulfilled [10, 9]: $\{\pi_i \sim d_i / \sum_i d_i\}$. Consequently, each edge in the network, $\{e_{i \rightarrow j}\}$, has a probability $\{\pi_{i \rightarrow j} = \pi_i / d_i \sim 1 / \sum_i d_i\}$ to be sampled, and the observed recruitment matrix from the RDS sample is an unbiased estimate of S .

However, when the network is directed, the inclusion probability for a node is no longer proportional to its degree, and the observed recruitment matrix from the sample will be representative only if individuals of the same group have similar edge formations, i.e., the personal recruitment matrix, $\{S_{XY}^i\}$, is the same for individuals in group X . Then, the observed raw recruitment matrix could be an appropriate estimate of S_{XY} .

A more general way is to develop a Hansen-Hurwitz type estimator for S_{XY} using the edges' inclusion probabilities $(\{\pi_{i \rightarrow j} = \pi_i / d_i^{out}\})$:

$$\hat{s}_{XY} = \frac{\sum_{i \rightarrow j, i \in X, j \in Y} \frac{d_j^{out}}{\pi_i}}{\sum_{i \rightarrow j, i \in X} \frac{d_j^{out}}{\pi_i}}, \quad (18)$$

where X and Y are the set of nodes with corresponding properties in the sample. Since $\{\pi_i\}$ is usually not known when knowledge about the structure of the network is incomplete, we might use the mean field approach to approximate $\{\pi_i\}$ with the average inclusion probability for nodes within group X :

$$\hat{s}_{XY} = \frac{\sum_{i \rightarrow j, i \in X, j \in Y} \frac{d_j^{out}}{\pi_X}}{\sum_{i \rightarrow j, i \in X} \frac{d_j^{out}}{\pi_X}} = \frac{\sum_{i \rightarrow j, i \in X, j \in Y} d_j^{out}}{\sum_{i \rightarrow j, i \in X} d_j^{out}}. \quad (19)$$

Frankly, both using the observed recruitment matrix from the sample, and approximating as described by (19), are brutal methods for estimating S_{XY} . We have tried both in the SH_{in} estimator; however, it turns out that the adjustment for \hat{s} made by (19) always generates larger error and bias; we thus only provide the discussions here and choose not to show any results in the paper.

Appendix C: Supporting figures

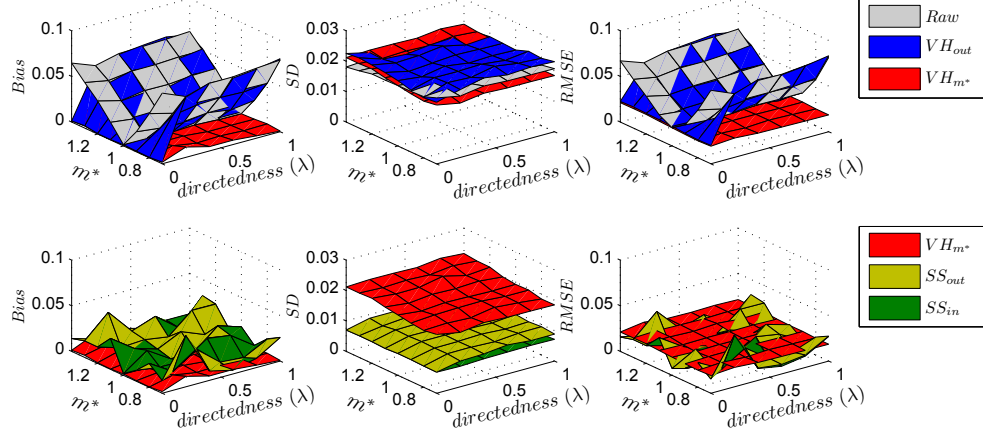


Figure 13: Bias, Standard Deviation, and Root Mean Square Error of RDS estimators on Net1. Sampling without replacement, number of seeds=6, coupons=2, sample size=500.

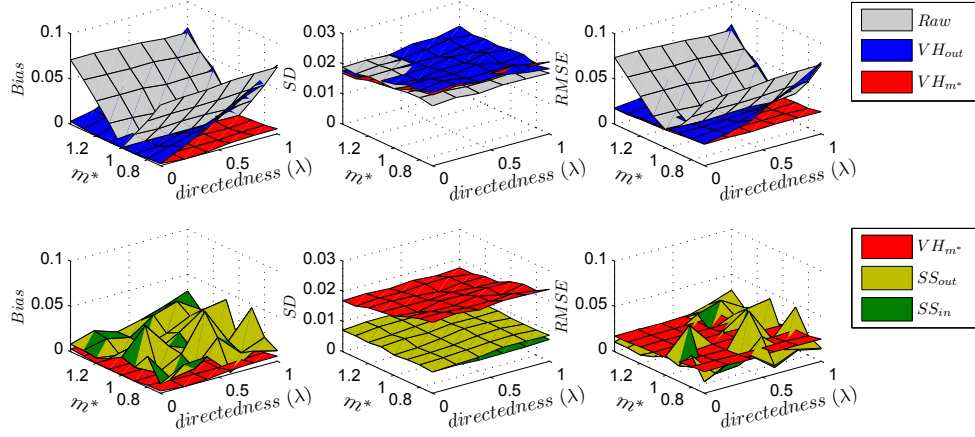


Figure 14: Bias, Standard Deviation, and Root Mean Square Error of RDS estimators on Net2, homophily $h_A = 0$. Sampling without replacement, number of seeds=6, coupons=2, sample size=500.

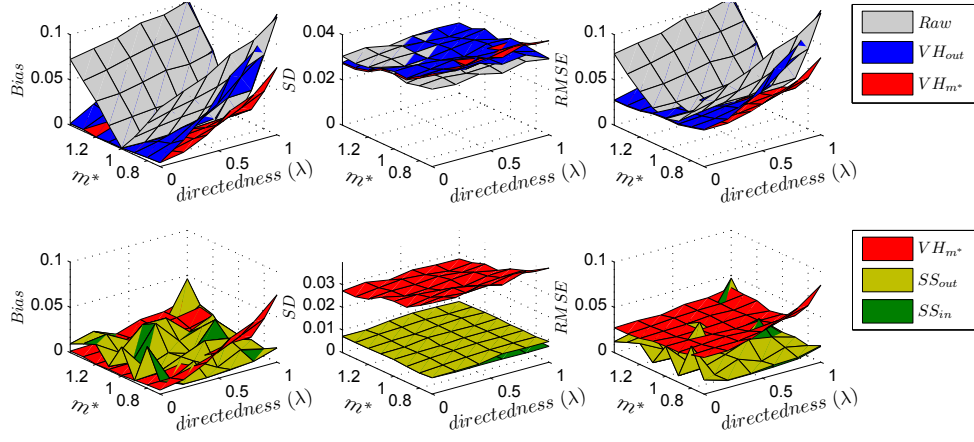


Figure 15: Bias, Standard Deviation, and Root Mean Square Error of RDS estimators on Net2, homophily $h_A = 0.4$. Sampling without replacement, number of seeds=6, coupons=2, sample size=500.

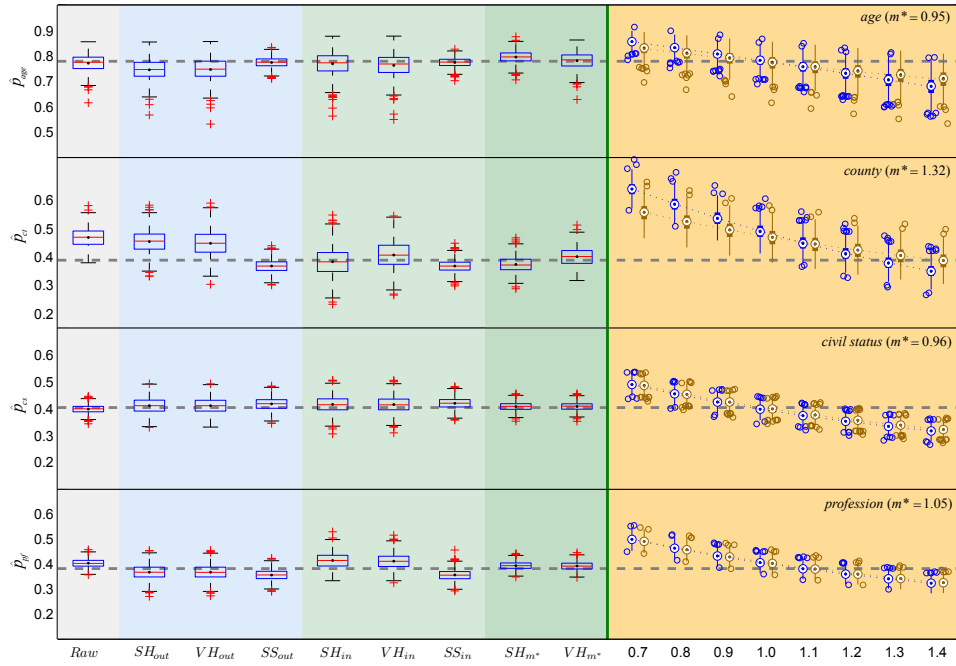


Figure 16: RDS on MSM network. The right panel shows sensitivity analysis of \hat{p}_A^{VHm} (brown) and \hat{p}_A^{SHm} (blue) with m varying from 0.7 to 1.4, plots are horizontally shifted a few points to avoid overlapping. Sampling with replacement, number of seeds=6, number of coupons=2, sample size=1000.

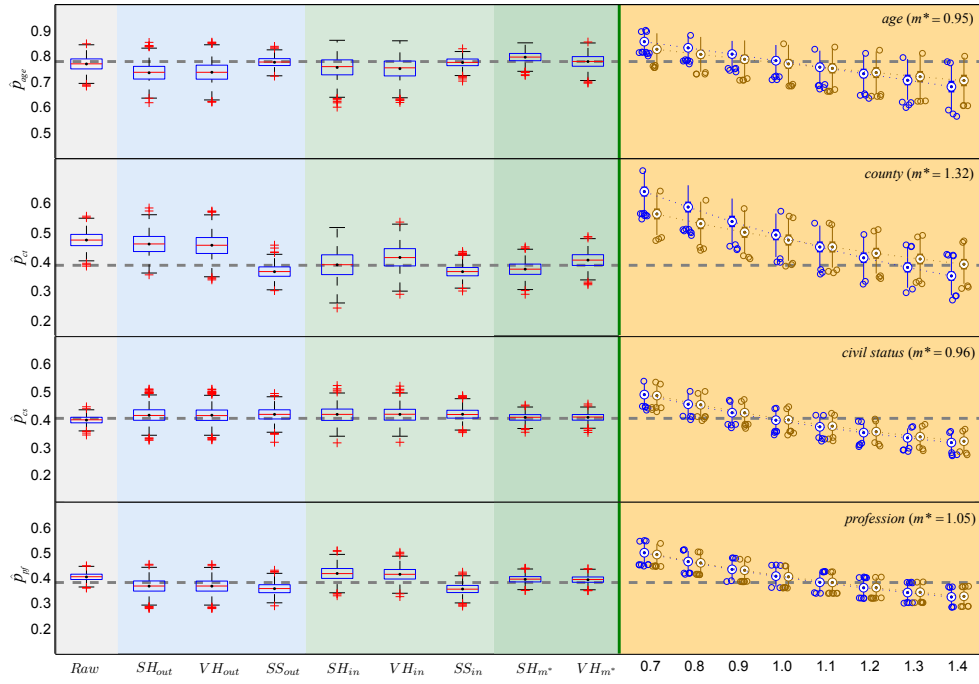


Figure 17: RDS on Net3 with indegree correlation $\gamma = 0.4$. The right panel shows sensitivity analysis of \hat{p}_A^{VHm} (brown) and \hat{p}_A^{SHm} (blue) with m varying from 0.7 to 1.4, plots are horizontally shifted a few points to avoid overlapping. Sampling without replacement, number of seeds=6, number of coupons=2, sample size=1000.

	Φ_{Net1}^{90}								Φ_{Net1}^{95}							
	0.7	0.8	0.9	m^* 1.0	1.1	1.2	1.3	1.4	0.7	0.8	0.9	m^* 1.0	1.1	1.2	1.3	1.4
$directedness(\lambda)$																
0.0	.05 (.90)	.30 (.90)	.75 (.91)	.91 (.90)	.77 (.92)	.48 (.90)	.22 (.90)	.09 (.91)	.07 (.95)	.41 (.95)	.84 (.95)	.96 (.95)	.86 (.96)	.58 (.95)	.30 (.95)	.13 (.96)
0.2	.20 (.85)	.53 (.88)	.83 (.90)	.90 (.89)	.84 (.88)	.69 (.90)	.48 (.87)	.32 (.88)	.32 (.91)	.66 (.94)	.90 (.94)	.95 (.95)	.91 (.93)	.80 (.94)	.60 (.91)	.44 (.93)
0.4	.08 (.89)	.42 (.90)	.79 (.88)	.90 (.91)	.80 (.92)	.59 (.89)	.35 (.89)	.13 (.89)	.14 (.94)	.53 (.96)	.89 (.94)	.95 (.95)	.88 (.96)	.71 (.95)	.48 (.94)	.22 (.94)
0.6	.07 (.91)	.37 (.90)	.75 (.92)	.89 (.94)	.78 (.91)	.56 (.90)	.27 (.88)	.17 (.90)	.14 (.95)	.52 (.95)	.86 (.96)	.95 (.96)	.85 (.94)	.67 (.95)	.37 (.94)	.26 (.95)
0.8	.05 (.90)	.38 (.91)	.78 (.89)	.92 (.91)	.76 (.89)	.51 (.91)	.29 (.91)	.11 (.92)	.10 (.96)	.51 (.95)	.87 (.94)	.95 (.96)	.84 (.94)	.63 (.96)	.40 (.96)	.17 (.94)
1.0	.03 (.92)	.31 (.91)	.72 (.91)	.91 (.90)	.76 (.90)	.48 (.90)	.23 (.91)	.09 (.91)	.07 (.96)	.44 (.97)	.82 (.97)	.95 (.95)	.84 (.96)	.60 (.94)	.34 (.96)	.15 (.94)
	$\Phi_{Net2,h_A=0}^{90}$								$\Phi_{Net2,h_A=0}^{95}$							
	0.7	0.8	0.9	m^* 1.0	1.1	1.2	1.3	1.4	0.7	0.8	0.9	m^* 1.0	1.1	1.2	1.3	1.4
$directedness(\lambda)$																
0.0	.16 (.91)	.44 (.88)	.75 (.89)	.91 (.90)	.73 (.91)	.37 (.90)	.11 (.90)	.01 (.88)	.23 (.95)	.55 (.95)	.83 (.95)	.96 (.96)	.82 (.96)	.49 (.94)	.18 (.95)	.02 (.94)
0.2	.17 (.89)	.46 (.91)	.74 (.90)	.90 (.91)	.77 (.89)	.42 (.90)	.14 (.92)	.02 (.91)	.26 (.94)	.59 (.95)	.85 (.95)	.95 (.96)	.86 (.95)	.55 (.95)	.22 (.97)	.04 (.96)
0.4	.14 (.90)	.42 (.91)	.73 (.90)	.90 (.90)	.74 (.91)	.43 (.90)	.17 (.92)	.02 (.91)	.22 (.94)	.55 (.96)	.82 (.95)	.96 (.95)	.85 (.95)	.57 (.95)	.26 (.96)	.05 (.96)
0.6	.11 (.88)	.43 (.90)	.77 (.91)	.91 (.90)	.76 (.90)	.48 (.91)	.21 (.89)	.04 (.88)	.18 (.94)	.55 (.94)	.84 (.96)	.96 (.95)	.85 (.95)	.64 (.96)	.32 (.95)	.08 (.93)
0.8	.08 (.89)	.38 (.92)	.73 (.89)	.91 (.91)	.75 (.91)	.50 (.90)	.21 (.90)	.06 (.86)	.15 (.94)	.50 (.96)	.83 (.95)	.96 (.96)	.84 (.95)	.62 (.96)	.31 (.95)	.11 (.92)
1.0	.06 (.90)	.37 (.90)	.68 (.92)	.91 (.93)	.76 (.90)	.48 (.89)	.20 (.89)	.07 (.88)	.12 (.95)	.48 (.95)	.78 (.96)	.95 (.97)	.86 (.96)	.60 (.95)	.31 (.94)	.13 (.94)
	$\Phi_{Net2,h_A=0.4}^{90}$								$\Phi_{Net2,h_A=0.4}^{95}$							
	0.7	0.8	0.9	m^* 1.0	1.1	1.2	1.3	1.4	0.7	0.8	0.9	m^* 1.0	1.1	1.2	1.3	1.4
$directedness(\lambda)$																
0.0	.23 (.90)	.55 (.89)	.83 (.91)	.96 (.92)	.82 (.91)	.49 (.91)	.18 (.91)	.02 (.92)	.80 (.95)	.89 (.94)	.93 (.96)	.96 (.96)	.94 (.96)	.88 (.95)	.79 (.96)	.61 (.96)
0.2	.26 (.89)	.59 (.89)	.85 (.89)	.95 (.92)	.86 (.89)	.55 (.91)	.22 (.90)	.04 (.91)	.63 (.94)	.79 (.94)	.91 (.94)	.96 (.95)	.93 (.95)	.84 (.95)	.67 (.95)	.51 (.95)
0.4	.22 (.81)	.55 (.86)	.82 (.89)	.96 (.90)	.85 (.90)	.57 (.90)	.26 (.90)	.05 (.89)	.46 (.86)	.70 (.92)	.89 (.94)	.96 (.95)	.90 (.95)	.78 (.95)	.56 (.95)	.36 (.94)
0.6	.18 (.64)	.55 (.85)	.84 (.89)	.96 (.91)	.85 (.92)	.64 (.88)	.32 (.87)	.08 (.84)	.24 (.76)	.64 (.91)	.86 (.94)	.96 (.96)	.89 (.96)	.69 (.94)	.42 (.93)	.23 (.90)
0.8	.15 (.53)	.50 (.75)	.83 (.85)	.96 (.92)	.84 (.91)	.62 (.83)	.31 (.85)	.11 (.80)	.12 (.65)	.46 (.83)	.82 (.92)	.95 (.96)	.86 (.96)	.61 (.91)	.37 (.91)	.14 (.89)
1.0	.12 (.25)	.48 (.59)	.78 (.84)	.95 (.90)	.86 (.91)	.60 (.85)	.31 (.80)	.13 (.77)	.02 (.35)	.26 (.71)	.79 (.92)	.96 (.95)	.89 (.95)	.55 (.92)	.26 (.89)	.07 (.86)

Figure 18: 90% and 95% bootstrap coverage probability of $\hat{p}_A^{VH_{out}}$ and $\hat{p}_A^{VH_{m^*}}$ (shown in brackets) on Net1 and Net2. Sampling without replacement, number of seeds=6, coupons=2, sample size=500.

References

- [1] UN Joint Programme on HIV/AIDS, Global report: Unaid report on the global aids epidemic 2010, Tech. rep. (2010).
- [2] M. Malekinejad, L. G. Johnston, C. Kendall, L. Kerr, M. R. Rifkin, G. W. Rutherford, Using respondent-driven sampling methodology for hiv biological and behavioral surveillance in international settings: A systematic review, *Aids and Behavior* 12 (4) (2008) S105–S130.
- [3] S. Goel, M. J. Salganik, Respondent-driven sampling as markov chain monte carlo, *Statistics in Medicine* 28 (17) (2009) 2202–2229.
- [4] E. Deaux, J. W. Callaghan, Key informant versus self-report estimates of health-risk behavior, *Evaluation Review* 9 (3) (1985) 365–368.
- [5] J. K. Watters, P. Biernacki, Targeted sampling: Options for the study of hidden populations, *Social Problems* 36 (4) (1989) 416–430.
- [6] B. H. Erickson, Some problems of inference from chain data, *Sociological Methodology* 10 (1979) 276–302.
- [7] D. D. Heckathorn, Respondent-driven sampling: A new approach to the study of hidden populations, *Social Problems* 44 (2) (1997) 174–199.
- [8] D. D. Heckathorn, Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples of hidden populations, *Social Problems* 49 (1) (2002) 11–34.
- [9] E. Volz, D. D. Heckathorn, Probability based estimation theory for respondent driven sampling, *Journal of Official Statistics* 24 (1) (2008) 79–97.
- [10] M. J. Salganik, D. D. Heckathorn, Sampling and estimation in hidden populations using respondent-driven sampling, *Sociological Methodology* 34 (2004) 193–239.
- [11] L. G. Johnston, M. Malekinejad, C. Kendall, I. M. Iuppa, G. W. Rutherford, Implementation challenges to using respondent-driven sampling methodology for hiv biological and behavioral surveillance: Field experiences in international settings, *Aids and Behavior* 12 (4) (2008) S131–S141.
- [12] S. Goel, M. J. Salganik, Assessing respondent-driven sampling, *Proceedings of the National Academy of Sciences of the United States of America* 107 (15) (2010) 6743–6747.
- [13] K. J. Gile, M. S. Handcock, Respondent-driven sampling: An assessment of current methodology, *Sociological Methodology* 40 (2010) 285–327.

- [14] A. Tomas, K. J. Gile, The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling, *Electronic Journal of Statistics* 5 (2011) 899–934.
- [15] K. J. Gile, Improved inference for respondent-driven sampling data with application to hiv prevalence estimation, *Journal of the American Statistical Association* 106 (493) (2011) 135–146.
- [16] S. J. South, D. L. Haynie, Friendship networks of mobile adolescents, *Social Forces* 83 (1) (2004) 315–350.
- [17] W. L. Wallace, *Student culture: Social structure and continuity in a liberal arts college*, Aldine Publishing Company, Chicago, 1966.
- [18] S. L. Feld, W. C. Carter, Detecting measurement bias in respondent reports of personal networks, *Social Networks* 24 (4) (2002) 365–383.
- [19] D. M. Paquette, J. Bryant, J. De Wit, Use of respondent-driven sampling to enhance understanding of injecting networks: A study of people who inject drugs in sydney, australia, *International Journal of Drug Policy* 22 (4) (2011) 267–273.
- [20] D. Abramovitz, E. M. Volz, S. A. Strathdee, T. L. Patterson, A. Vera, S. D. W. Frost, P. ElCuete, Using respondent-driven sampling in a hidden population at risk of hiv infection: Who do hiv-positive recruiters recruit?, *Sexually Transmitted Diseases* 36 (12) (2009) 750–756.
- [21] X. Y. Ma, Q. Y. Zhang, X. He, W. D. Sun, H. Yue, S. Chen, H. F. Raymond, Y. Li, M. Xu, H. Du, W. McFarland, Trends in prevalence of hiv, syphilis, hepatitis c, hepatitis b, and sexual risk behavior among men who have sex with men - results of 3 consecutive respondent-driven sampling surveys in beijing, 2004 through 2006, *J AIDS-Journal of Acquired Immune Deficiency Syndromes* 45 (5) (2007) 581–587.
- [22] M. Y. Iguchi, A. J. Ober, S. H. Berry, T. Fain, D. D. Heckathorn, P. M. Gorbach, R. Heimer, A. Kozlov, L. J. Ouellet, S. Shoptaw, W. A. Zule, Simultaneous recruitment of drug users and men who have sex with men in the united states and russia using respondent-driven sampling: Sampling methods and implications, *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 86 (2009) S5–S31.
- [23] X. Lu, L. Bengtsson, T. Britton, M. Camitz, B. J. Kim, A. Thorson, F. Liljeros, The sensitivity of respondent-driven sampling, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175 (1) (2012) 191–216.
- [24] W. K. Hastings, Monte-carlo sampling methods using markov chains and their applications, *Biometrika* 57 (1) (1970) 97–109.
- [25] S. Fortunato, M. Boguñá, A. Flammini, F. Menczer, Algorithms and models for the web-graph, 2008, Ch. Approximating PageRank from In-Degree, pp. 59–71.
- [26] M. E. J. Newman, Assortative mixing in networks, *Physical Review Letters* 89 (20) (2002) 208701.
- [27] M. Morris, M. Kretzschmar, Concurrent partnerships and transmission dynamic in networks, *Social Networks* 17 (3-4) (1995) 299–318.
- [28] A. Rapoport, A probabilistic approach to networks, *Social Networks* 2 (1980) 1–18.
- [29] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks, *Annual Review of Sociology* 27 (2001) 415–444.
- [30] D. Rybski, S. Buldyrev, S. Havlin, F. Liljeros, H. Makse, Scaling laws of human interaction activity, *Proceedings of the National Academy of Sciences of the United States of America* 106 (31) (2009) 12640–12645.
- [31] M. J. Salganik, Variance estimation, design effects, and sample size calculations for respondent-driven sampling, *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 83 (6) (2006) I98–I112.
- [32] A. S. Abdul-Quader, D. D. Heckathorn, C. McKnight, H. Bramson, C. Nemeth, K. Sabin, K. Gallagher, D. C. Des Jarlais, Effectiveness of respondent-driven sampling for recruiting drug users in new york city: Findings from a pilot study, *Journal of Urban Health-Bulletin of the New York Academy of Medicine* 83 (3) (2006) 459–476.
- [33] C. Wejnert, D. D. Heckathorn, Web-based network sampling - efficiency and efficacy of respondent-driven sampling for online research, *Sociological Methods & Research* 37 (1) (2008) 105–134.
- [34] E. T. O'Neill, P. D. McClain, B. F. Lavoie, A methodology for sampling the world wide web web, *Journal of Library Administration* 34 (3/4) (2001) 279–291.
- [35] C. Snelson, Sampling the web: The development of a custom search tool for research, *Library and Information Science Research Electronic Journal* 16 (1).
- [36] M. Gjoka, M. Kurant, C. Butts, A. Markopoulou, Walking in facebook: A case study of unbiased sampling of osns, in: *INFOCOM, 2010 Proceedings IEEE*, 2010, pp. 1–9.
- [37] R. Xulvi-Brunet, I. M. Sokolov, Reshuffling scale-free networks: From random to assortative, *Physical Review E* 70 (6) (2004) 066102.